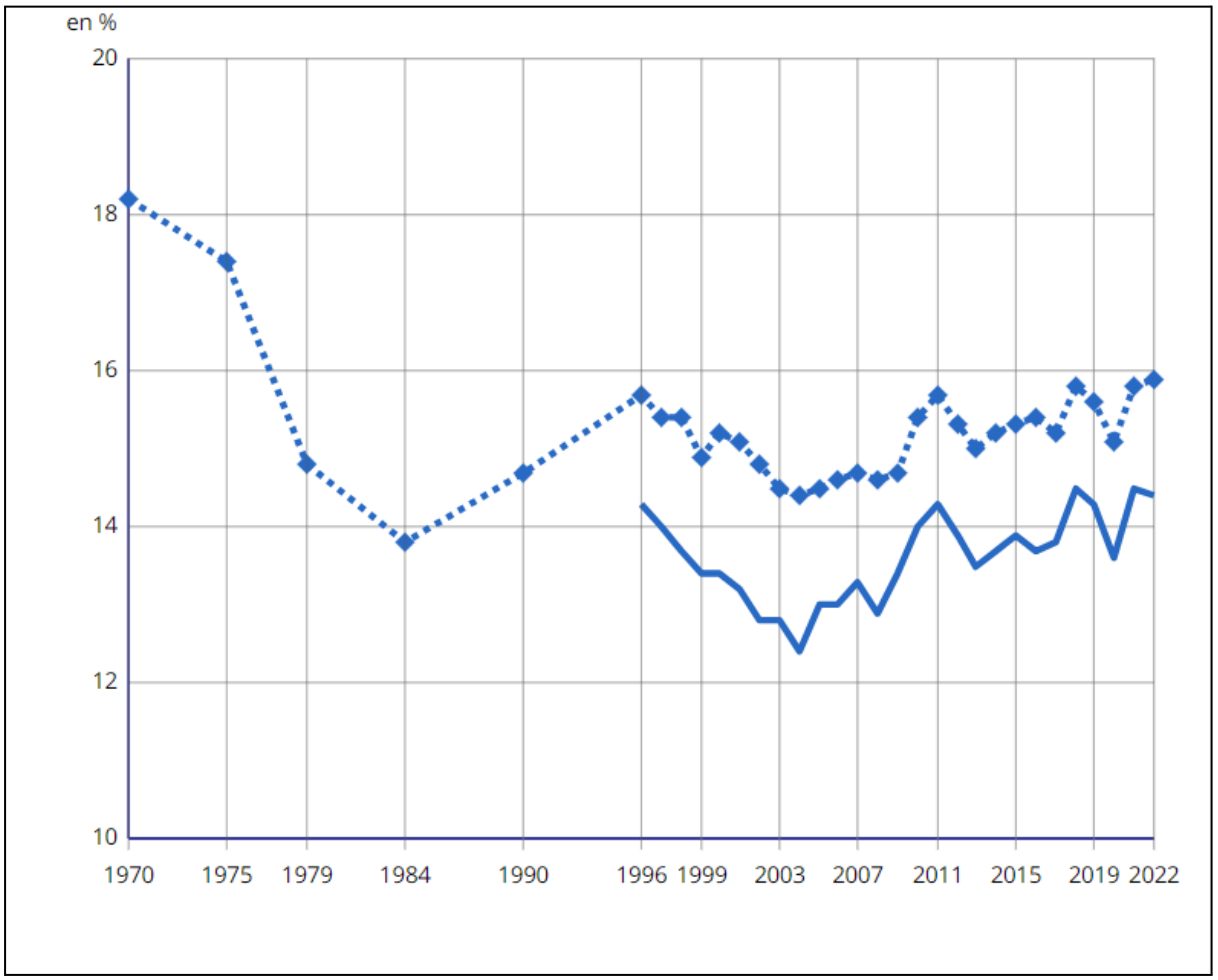
Léo GAUTHIER

Projet de l'UE Analyse de données

Étude sur le taux de pauvreté en France en 2020 : Quelles sont les différences entre les départements métropolitains et d'outre-mer ?

La problématique et la base de donnée utilisée

En France, bien que nous soyons un pays développé, le taux de pauvreté ne recule plus, mais il augmente même. La précarité est un sujet d'actualité, un enjeu moderne qui touche une part grandissante de la population française. En tant qu'étudiant, on la côtoie au quotidien. C'est dans un objectif de mieux la comprendre, et de l'analyser que j'ai choisi ce sujet.

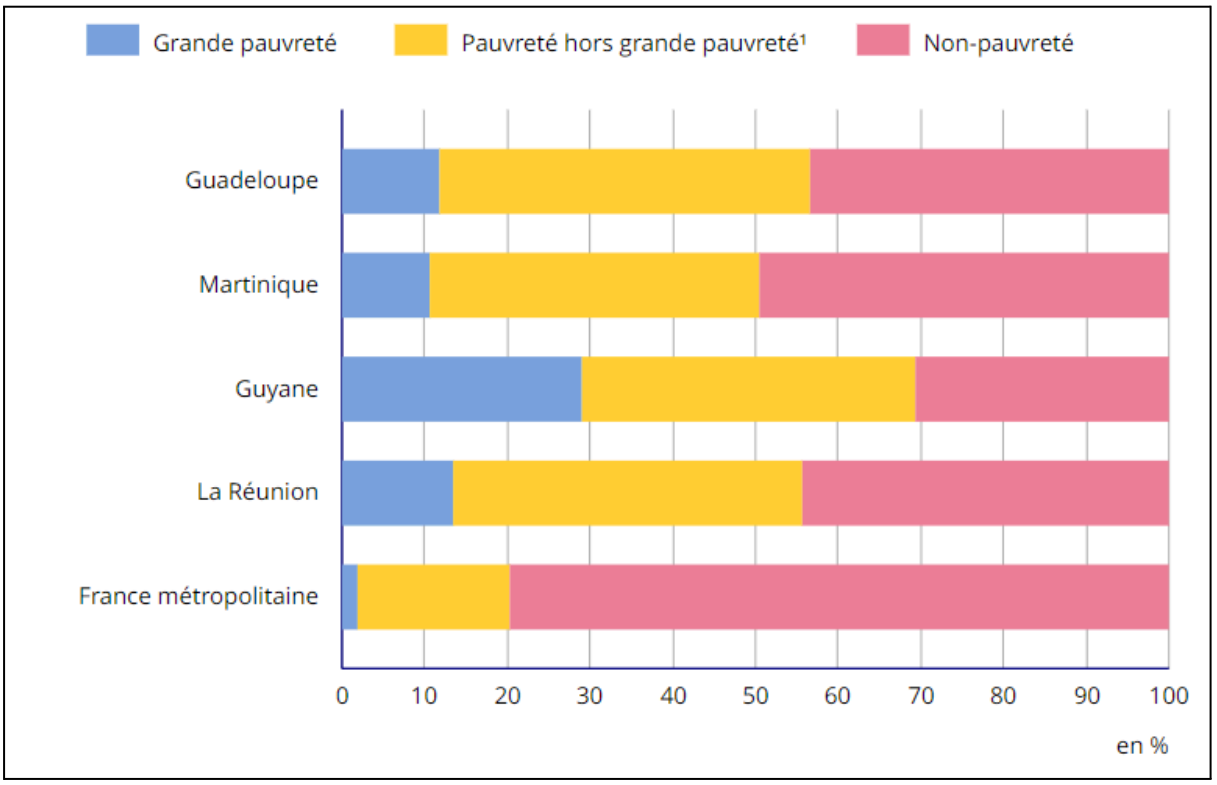


Taux de pauvreté depuis 1970[1]

[1] Source : https://www.insee.fr/fr/statistiques/5759045#graphique-figure4

Un sujet encore plus d'actualité est le fait que la pauvreté touche 5 à 10 fois plus[2] les départements d'Outre-mer que la France métropolitaine.

Ce projet vise donc à identifier et à analyser les sources des écarts qui engendrent cette différence.



Répartition de la population selon le taux de pauvreté en 2018[3]

Le jeu de données avec lequel j'ai choisi de travailler est une base que j'ai construite à partir de différentes sources et bases de données déjà existantes. C'est un jeu de données avec comme individus d'études les départements français. Elle recoupe 100 individus et 15 variables, qui sont :

- Le taux de pauvreté en 2020, en pourcentage
- Le niveau de vie médian (le revenu disponible du ménage rapporté au nombre d'unités de consommation (UC[4]) du ménage), en euro
- Le taux de chômage, en pourcentage
- La part des diplômés du supérieur parmis les 25-34 ans, en pourcentage
- Le salaire mensuel net moyen, en euro
- Les dépenses brutes en RSA, en euro
- Le nombre de ménages fiscaux (un regroupement de foyers fiscaux dans un même logement)

[2] Source : https://www.insee.fr/fr/statistiques/6459395#onglet-1
[3] Source : https://www.insee.fr/fr/statistiques/6459395#onglet-2
[4] 1 UC pour le premier adulte du ménage, 0,5 UC pour les autres personnes de 14 ans et plus, 0,3 UC pour les enfants de moins de 14 ans.

- La part de ces ménages fiscaux qui sont imposés, en pourcentage
- Le montant moyen de la taxe d'habitation par résidence principale, en euro
- Le taux de familles monoparentales, en pourcentage
- La part de la population française vivant dans une unité urbaine (villes de plus de 2000 habitants), en pourcentage
- Le taux d'activité chez les 25-54 ans, en pourcentage
- La population totale
- La densité de population au km2
- Les départements d'outre-mer (variable binaire, N si en métropole, O si en outre mer)

Ce sont principalement des données de 2020 et 2021 car ce sont les années où les documentations sont les plus complètes. Certaines données remontent à 2017 ou 2018
Il y a aussi bien des variables sur des données socio-démographiques (niveau de vie médian, etc…) que des variables plus économiques telle que le montant des dépenses RSA brutes. L'objectif était d'avoir un grand nombre de variables pour pouvoir décrire au mieux ce phénomène.
J'ai décidé de ne pas prendre en compte Mayotte car c'est un département avec des valeurs extrêmes pour un certain nombre de variables considérées, et cela risquerait de fausser l'analyse.

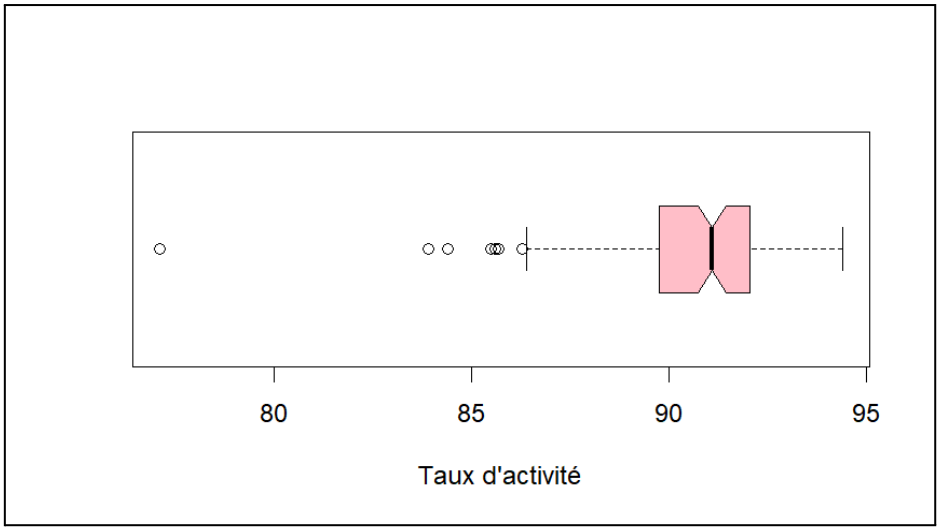
Les statistiques descriptives et des graphiques

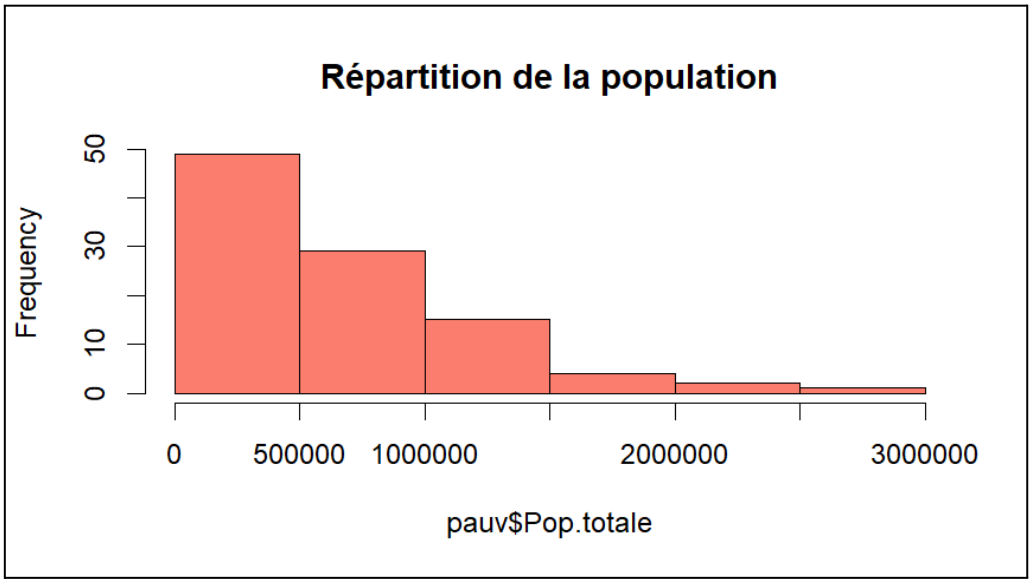
On calcule tout d'abord les statistiques descriptives :

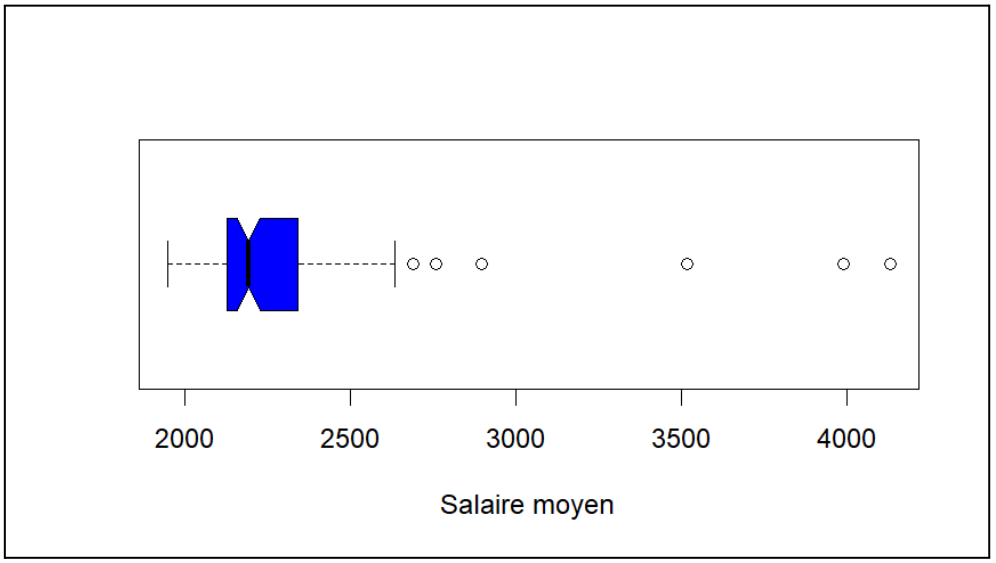
Variables	Moyenne	Ecart-type	Min	1er quartile	Médiane	3ème quartile	Max
Taux de pauvreté	15,26	5,72	8,90	12,28	14,30	16,20	53
Niveau de vie médian	22453	2 190,34	11300	21530	22225	23238	29730
Taux de chômage	8,092	2,15	4,8	6,8	7,650	8,092	17,5
Part des diplômés du sup.	40,11	8,22	21	34,95	38,30	43,08	81,10
Salaire mensuel moyen	2287	334,37	1948	2127	2192	2341	4131
Dépenses	116 526	117 625	7 450	41 007	74 525	145 596	681 172

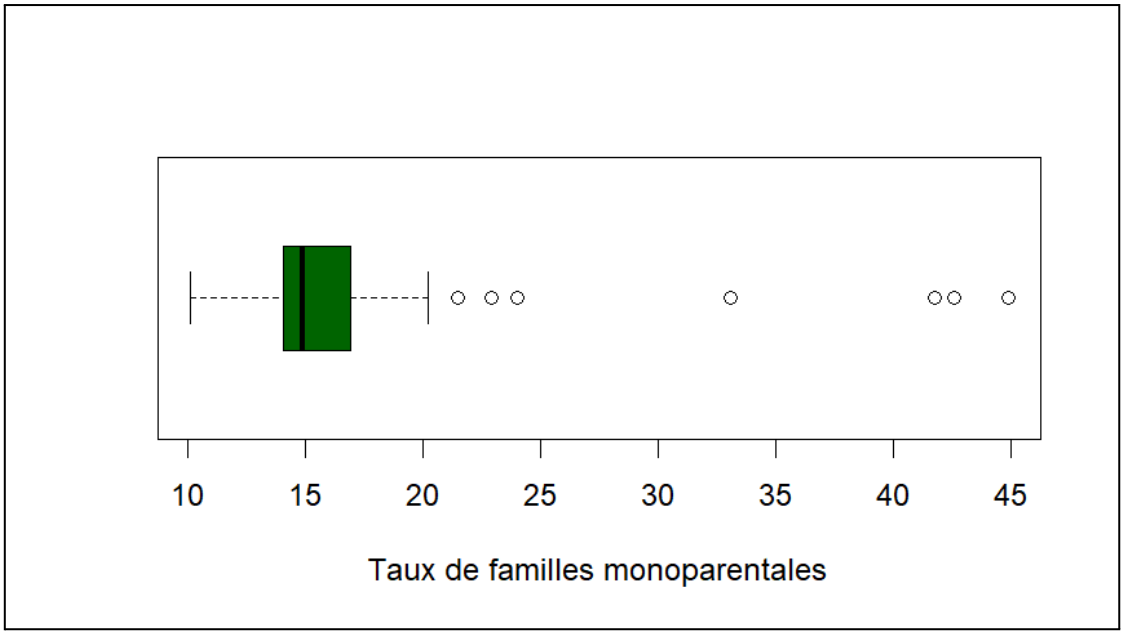
RSA brutes	357	000	402	334	242	499	702
Nombre de ménages fiscaux	290 016	217 325	34 028	130 609	231373	353 274	1 072 648
Part des ménages fiscaux imposés	50,03	6,96	25,10	46,15	49,15	53,38	70,40
Population totale	660 322	514 708	76 633	283 738	523 542	831 076	2 607 746
Taxe d'habitation moyenne par res. principale	208,7	99,63	88	140,2	179,8	233,8	605,3
Taux de famille monoparen-tales	16,36	5,57	10,10	14,07	14,90	16,93	44,90
Part de la population en unité urbaine	69,12	18,06	21,40	56,58	67,55	82,30	100
Taux d'activité chez les 25-54 ans	90,54	2,48	77,10	89,78	91,10	92,03	94,40
Densité de population	552,09	2 369,73	3,40	50,05	85,50	174,35	20 359,6

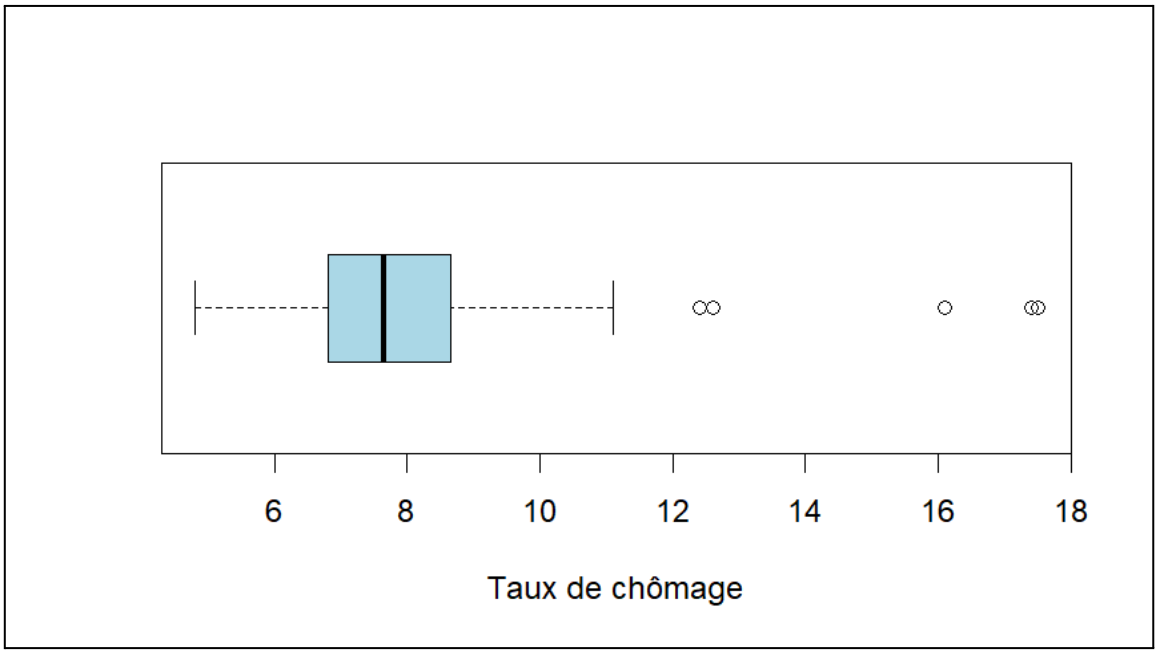
On réalise également quelques simples graphiques pour illustrer certaines variables :


80
85
90
95
Taux d'activité


Répartition de la population
Frequency
0
10
30
50
0
500000
1000000
2000000
3000000
pauv$Pop.totale


2000
2500
3000
3500
4000
Salaire moyen





L'analyse factorielle

Le jeu de données étant composé principalement de variables quantitatives, on décide naturellement de réaliser une analyse en composantes principales, une ACP. Cette analyse va nous permettre de faire des comparaisons entre les individus, et de pouvoir identifier les ressemblances. On aura aussi les corrélations entre variables.

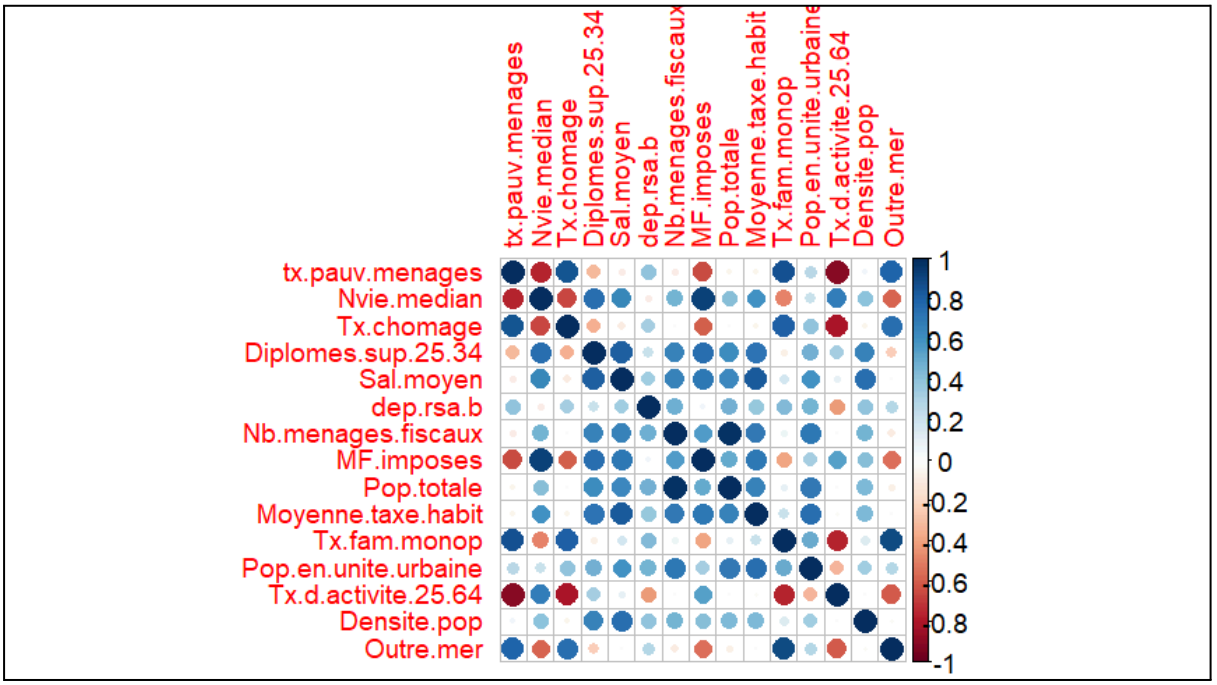
Or, il y a quand même une variable qui est qualitative dans le jeu de données : les départements d'outre-mer.

On va alors considérer en variables illustratives la densité de population au km2 et les départements d'outre mer, et en variables actives toutes les autres. Cela permettra de faciliter l'interprétation des axes, et de contextualiser les résultats. On sait également

que les variables illustratives pour une ACP peuvent être qualitatives, ce n'est pas un souci.

Puisque les données prennent des valeurs de taille extrêmement variables et qu'elles ont des unités de mesures différentes, on doit normaliser toutes les variables pour équilibrer leurs influences. La fonction PCA du package FactoMineR normalise de façon automatique les données si on ne précise pas l'inverse.

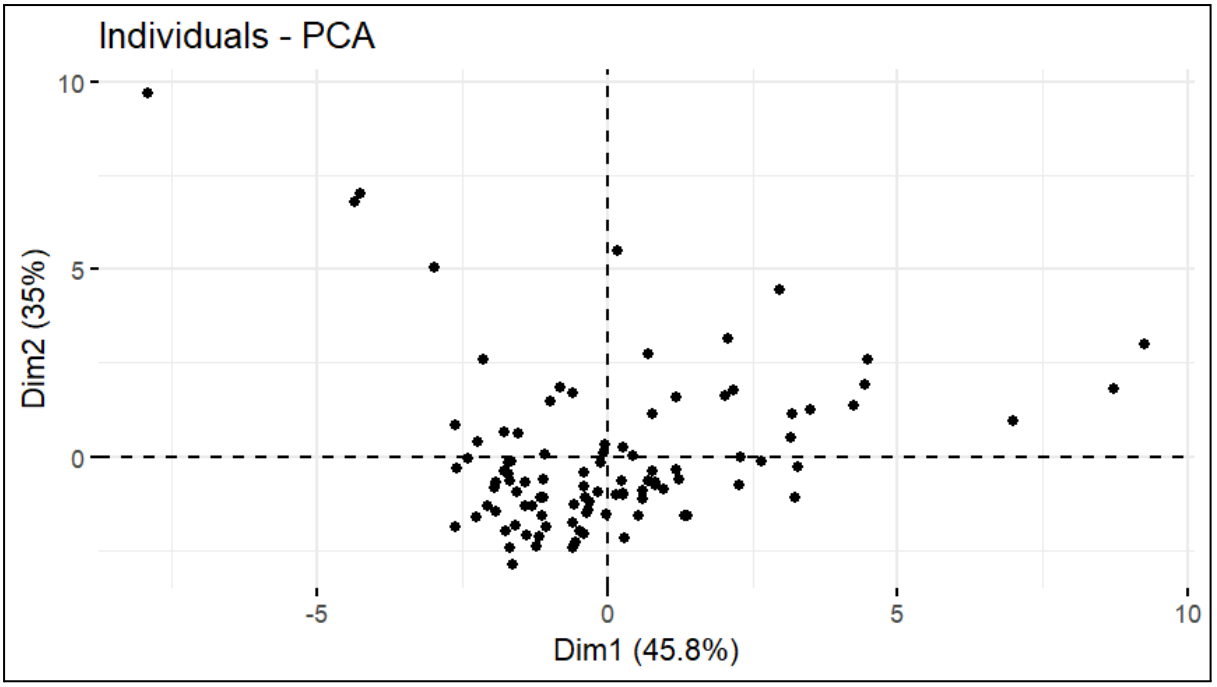
On regarde dans un premier temps la matrice de corrélations pour identifier les corrélations présentes entre variables:



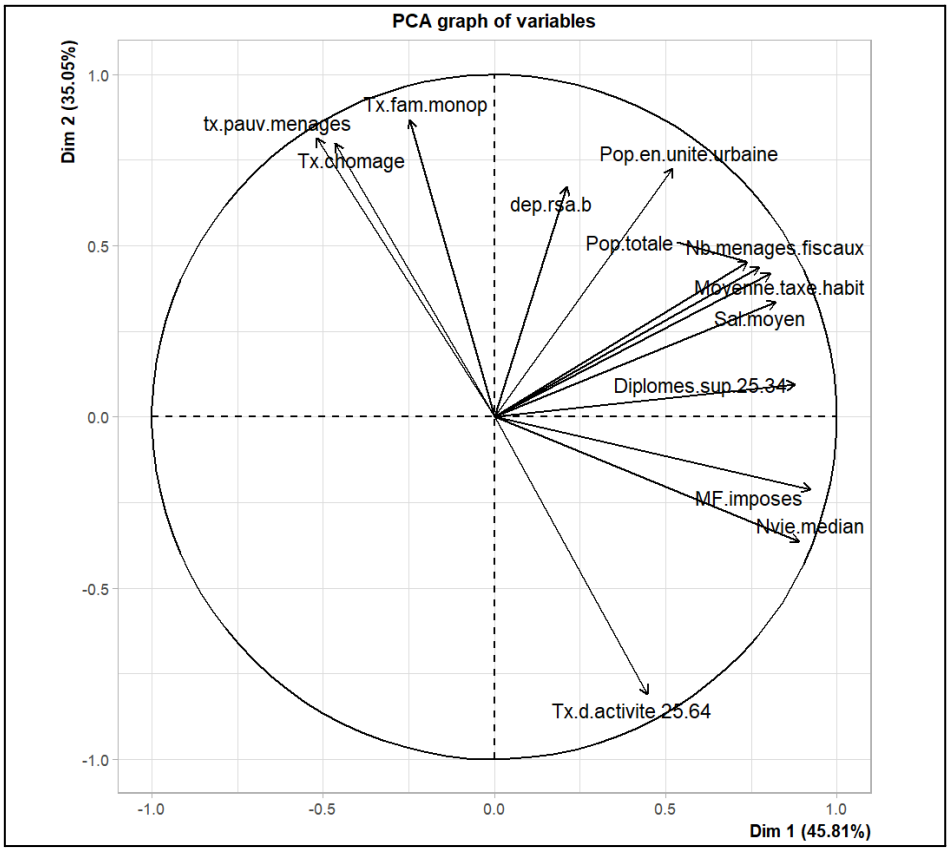
On remarque qu'il y a des corrélations positives importantes entre le taux de chômage et le taux de pauvreté des ménages, entre la population totale et le nombre de ménages fiscaux et entre la part des ménages fiscaux imposés et le niveau de vie médian. Aucune de ces corrélations n'est étonnante, ce sont toutes des variables qui sont étroitement liées.
Il y a également des variables qui entretiennent des corrélations négatives importantes, notamment le taux d'activité et le taux de pauvreté, et le taux de d'activité et le taux de chômage. Rien d'étonnant non plus.

Premiers résultats

On fait d'abord l'ACP sans prendre en compte les variables illustratives, et on les ajoutera après pour pouvoir comprendre leurs rôles. On a, pour le plan de projection des individus :

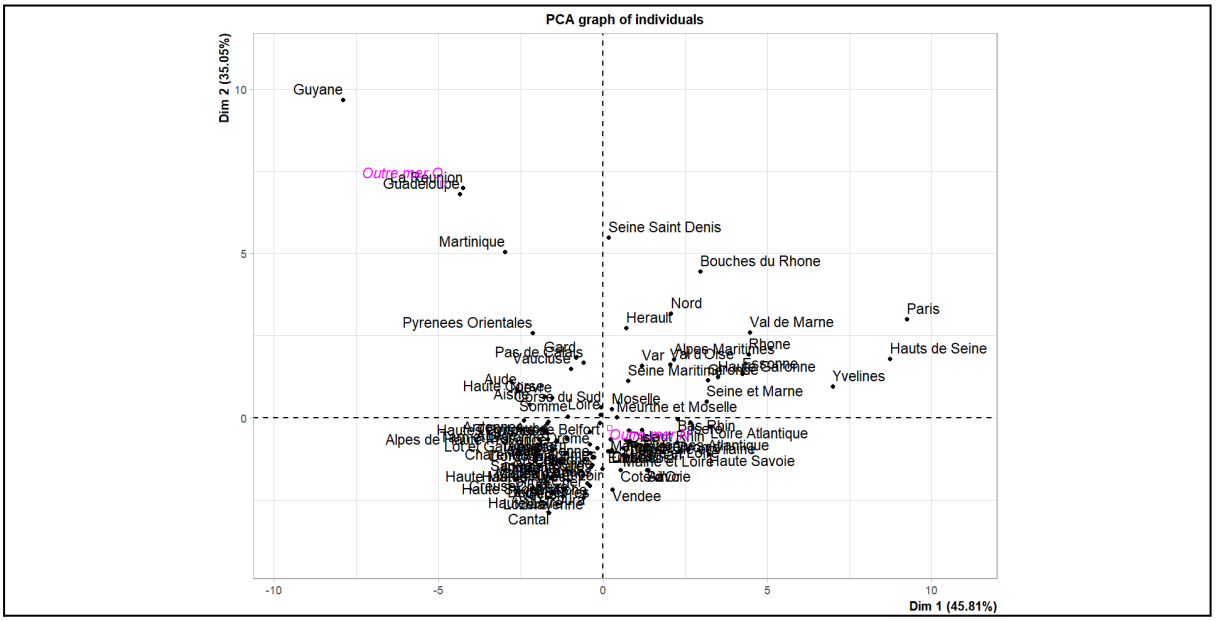


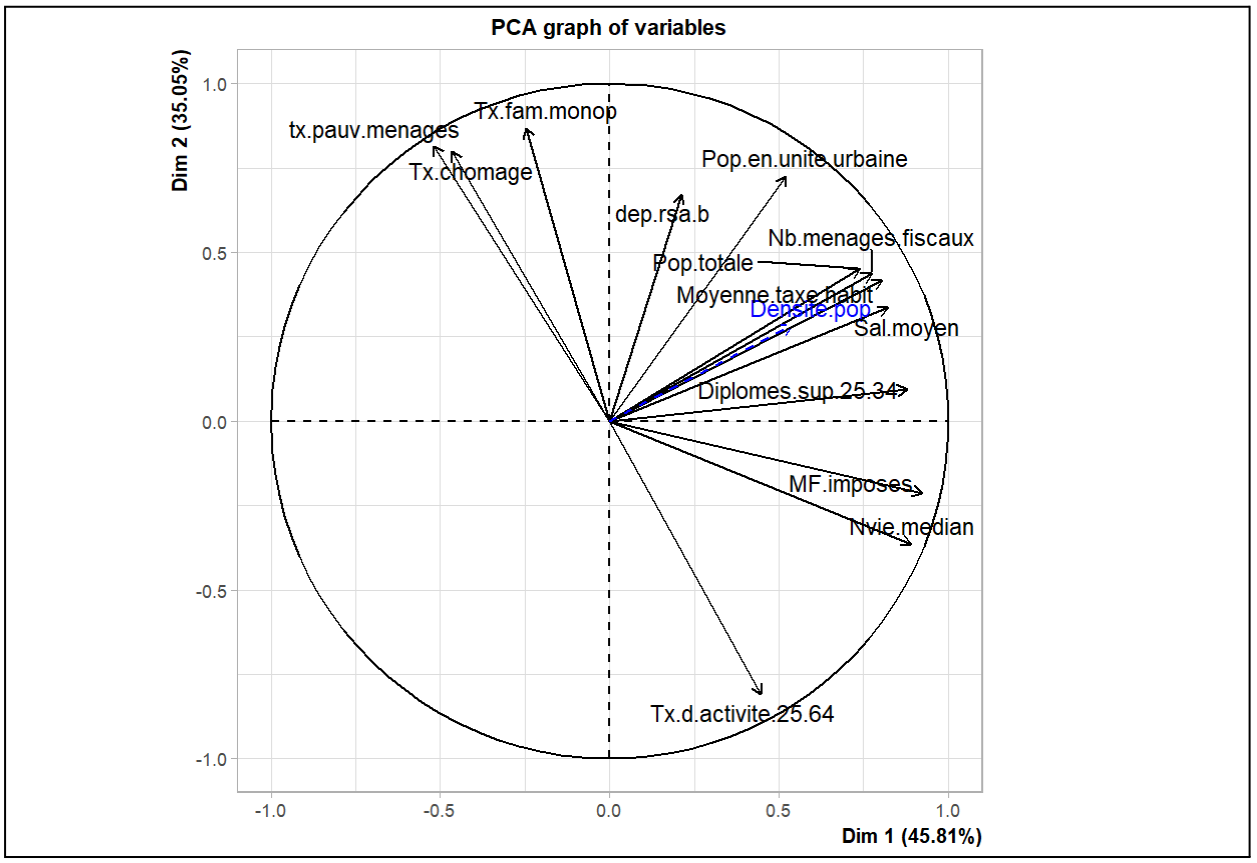
Et pour le plan de projection des variables :



Notre premier plan de projection récupère 80% de l'inertie totale, ce qui est un bon pourcentage.

On réalise ensuite l'ACP avec les variables illustratives, qui sera la base pour toutes nos analyses :





Dans le cadre de l'ACP normée, on a min(n,p) axes. Dans ce jeu de données, p=13 et n = 100, donc on a 13 axes.

On affiche les inerties de nos 13 axes :

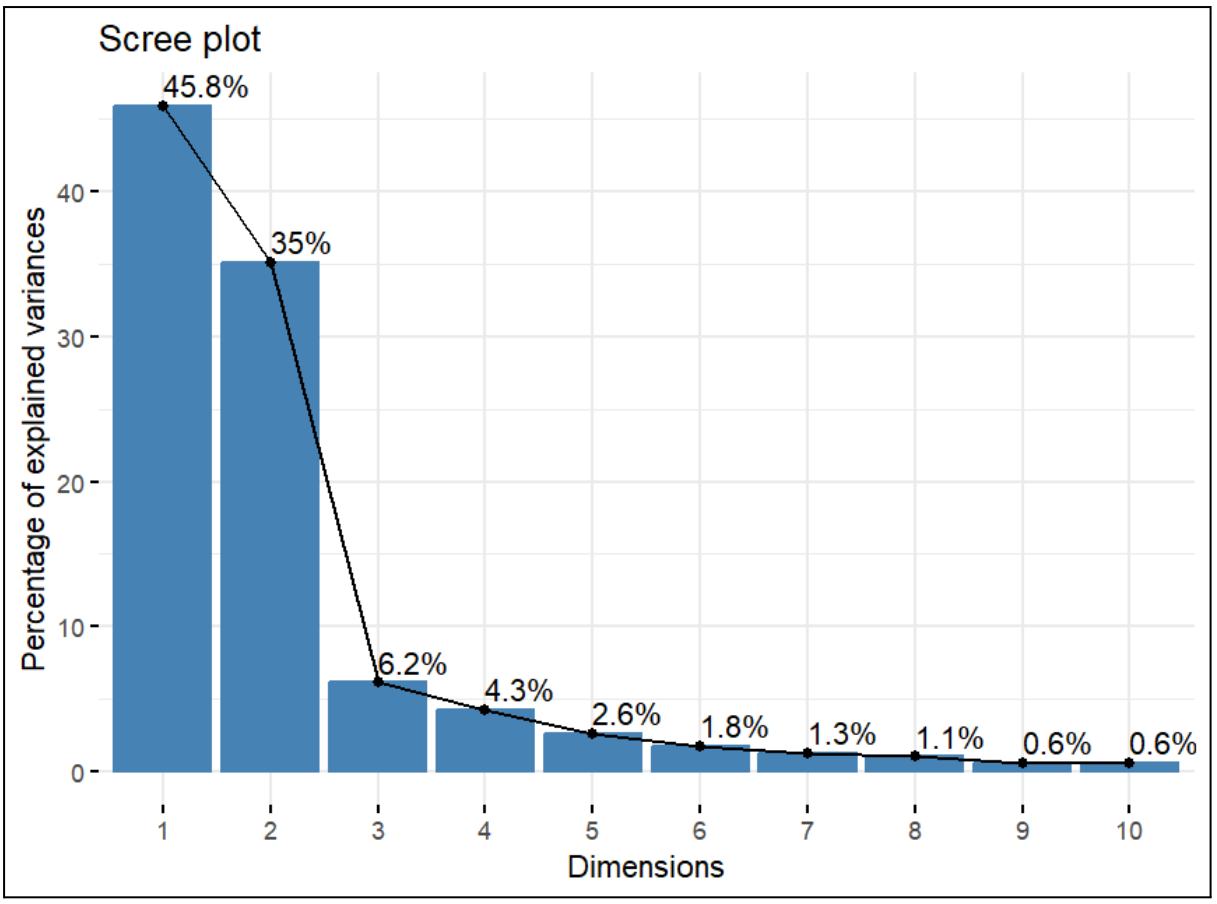
```
> pauv.acp1$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  5.95551140             45.8116262                            45.81163
comp 2  4.55645583             35.0496602                            80.86129
comp 3  0.80074309              6.1595623                            87.02085
comp 4  0.55588028              4.2760021                            91.29685
comp 5  0.34331987              2.6409221                            93.93777
comp 6  0.23051527              1.7731944                            95.71097
comp 7  0.16632699              1.2794384                            96.99041
comp 8  0.13866695              1.0666688                            98.05707
comp 9  0.08200907              0.6308390                            98.68791
comp 10 0.07774938              0.5980722                            99.28599
comp 11 0.05543684              0.4264373                            99.71242
comp 12 0.01988697              0.1529767                            99.86540
comp 13 0.01749805              0.1346004                           100.00000
```

```
> sum(pauv.acp1$eig[,1])
[1] 13
```

On a une inertie totale de notre ACP égale à 13. C'est normal car l'inertie totale, dans le cadre d'une ACP normée, est égale à p le nombre de variables actives. Ici, p=13. On a alors une inertie moyenne de 1 (car inertie moyenne = inertie totale / nombre d'axes).

Axes interprétables

Maintenant qu'on a regardé ces premiers résultats de l'ACP, on peut regarder le nombre de dimensions potentiellement intéressantes. On peut utiliser la méthode du bâton brisé ou le test de Catell, la recherche d'un coude sur l'éboulis des valeurs propres.
On a :



On a un coude assez clair : les deux premières dimensions sont de loin les plus intéressantes. On va cependant utiliser le critère du bâton brisé pour avoir des résultats plus précis :

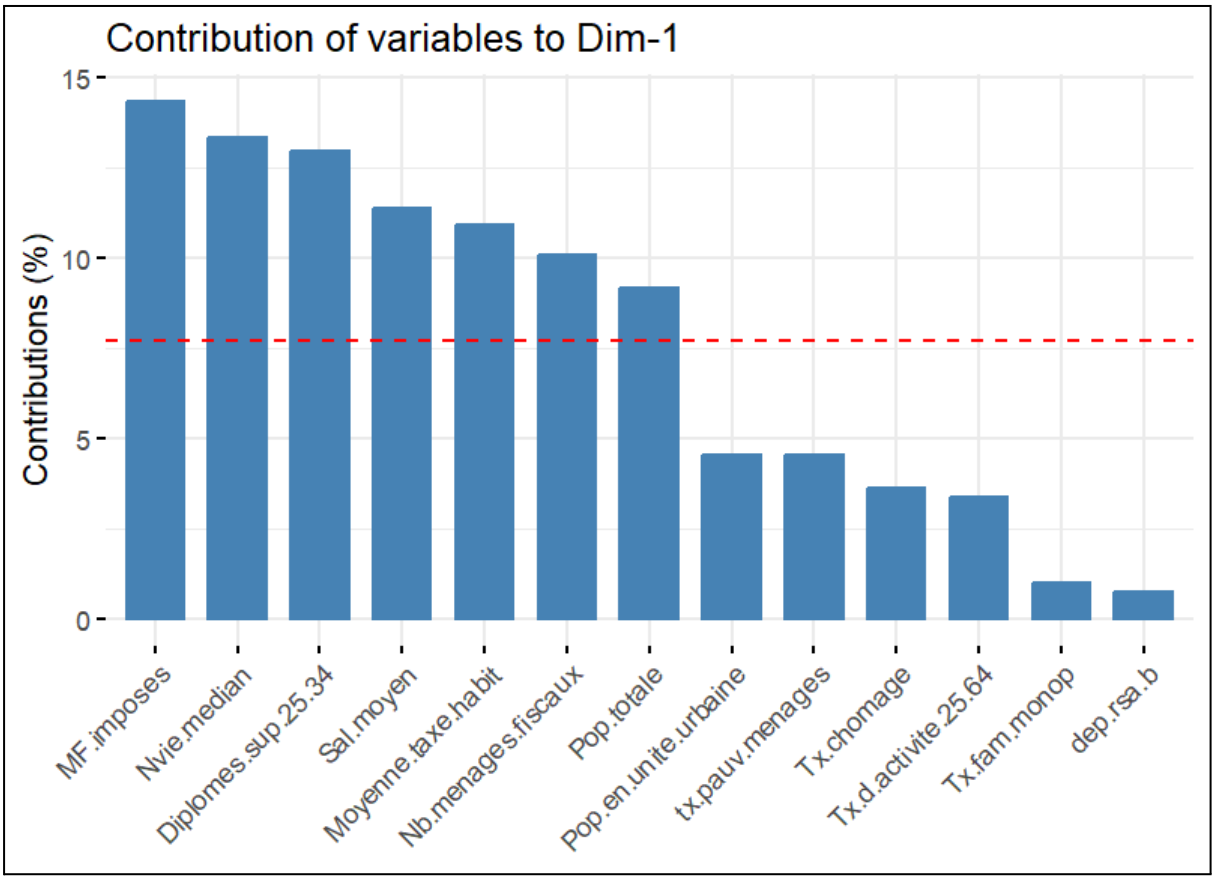
```
> brokenStick(1:10, 10)
 [1] 0.29289683 0.19289683 0.14289683 0.10956349 0.08456349 0.06456349
 [7] 0.04789683 0.03361111 0.02111111 0.01000000
```

La méthode du broken stick nous donne le pourcentage d'inertie minimal que doit avoir chaque axe pour pouvoir être intéressant à interpréter. On a donc la même conclusion qu'avec le test de Catell, les deux premiers axes factoriels sont les plus intéressants à interpréter.

Interprétation des axes

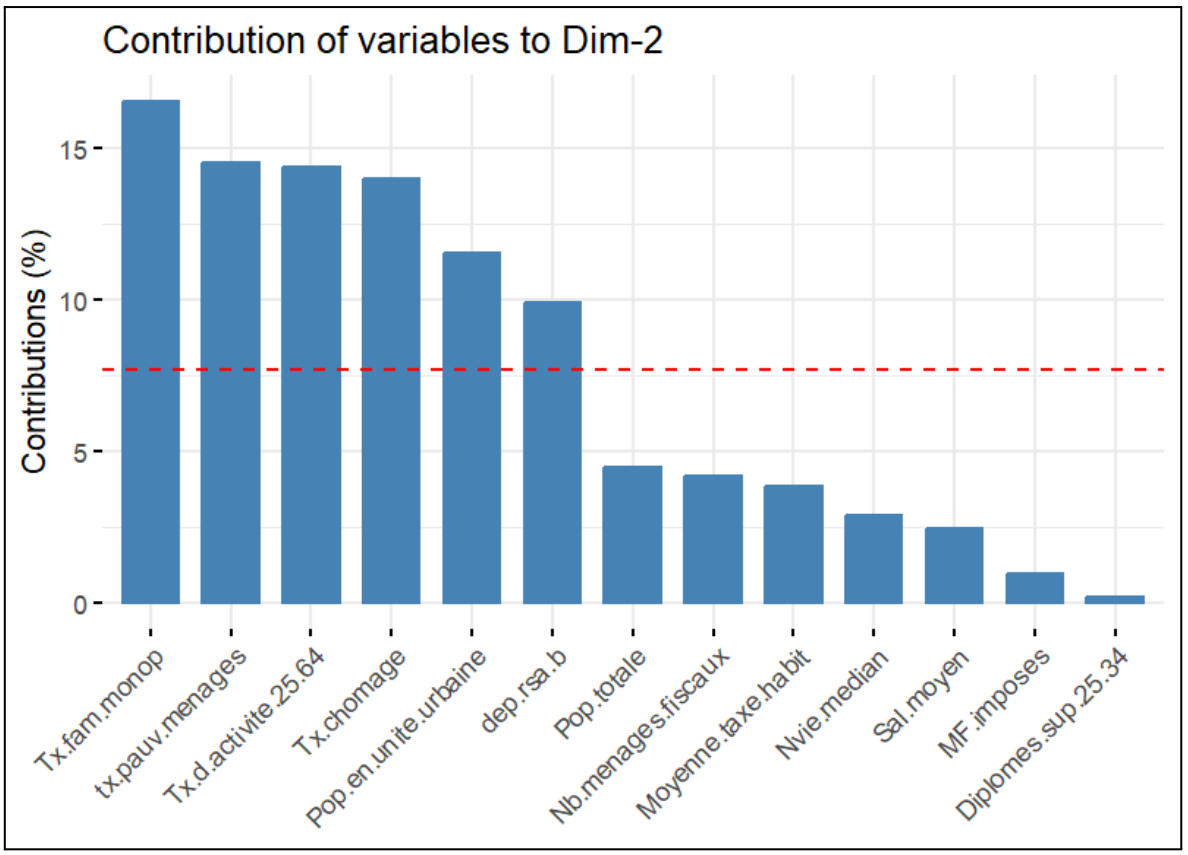
On peut alors interpréter les axes factoriels. Pour faire ceci, on va regarder quelles variables contribuent le plus à l'axe 1 puis à l'axe 2.

- **L'axe 1 :**



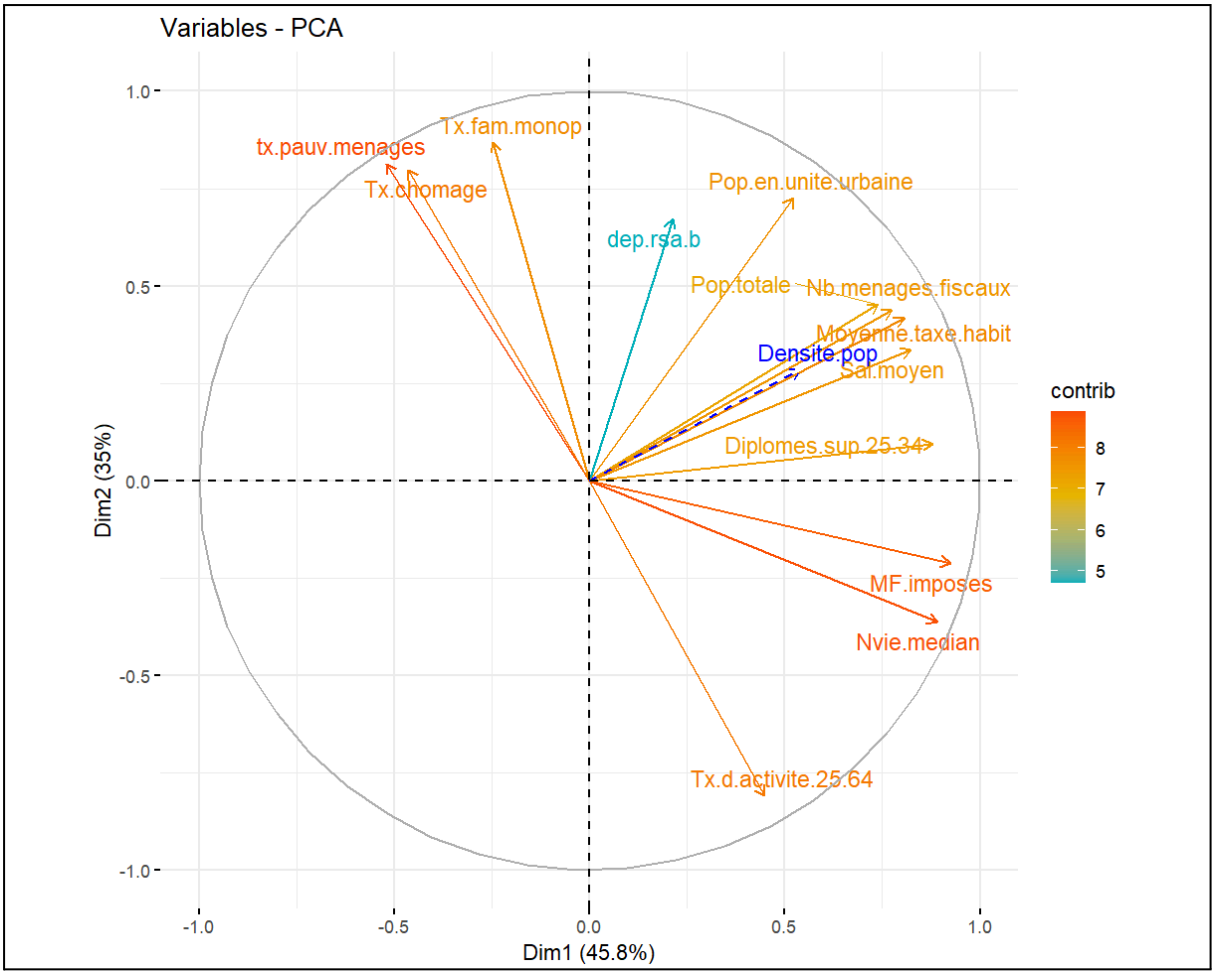
Les principales variables contributrices sont celles qui sont supérieures à la contribution moyenne des variables à l'axe 1, ici indiquée par la ligne de pointillés rouges.
L'axe 1 est alors une composante principale qui va plutôt capturer le niveau et qualité de vie du département, ainsi que les différences de population de celui-ci. Elles proposent un axe de richesse économique/prospérité de la population dans les départements. Ce sont plutôt des variables économiques qui seront contributrices à l'axe 1.

- **L'axe 2 :**

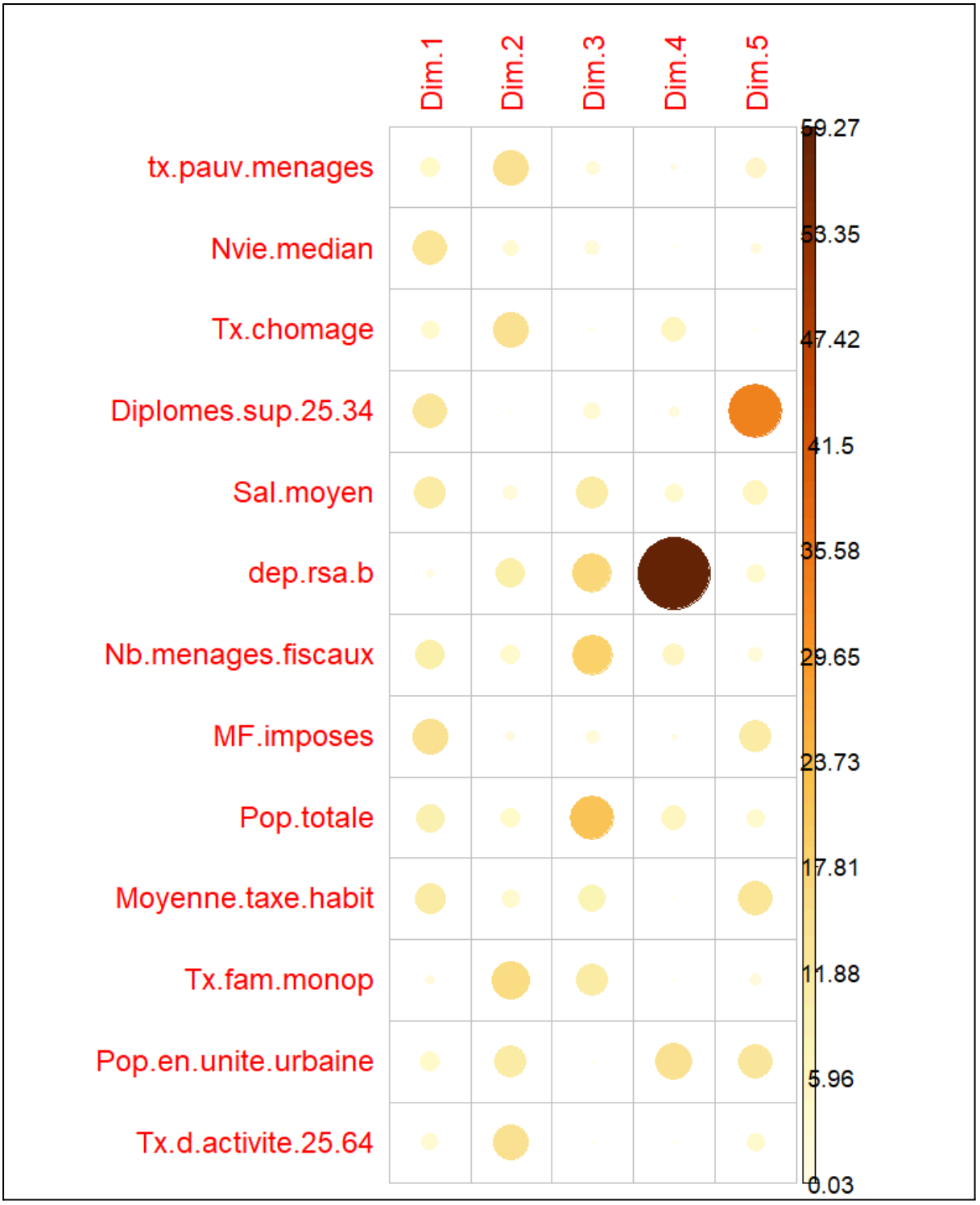


Pour la 2ème composante principale, ce sont des variables plus sociales, indicatrices d'une fragilité sociale qui vont être contributrices. Elles vont décrire les situations familiales, une certaine fragilité sociale.

Ce premier plan factoriel va alors opposer sur le premier axe des départements favorisés économiquement parlant aux départements moins riches, et sur le deuxième axe distinguer les territoires à fortes fragilités sociales des territoires plus stables.
On propose également un gradient en fonction des variables qui contribuent le plus au plan :



Et pour avoir un plus d'information sur l'ACP, on peut également s'intéresser aux corrélations des variables aux 5 premières composantes principales :

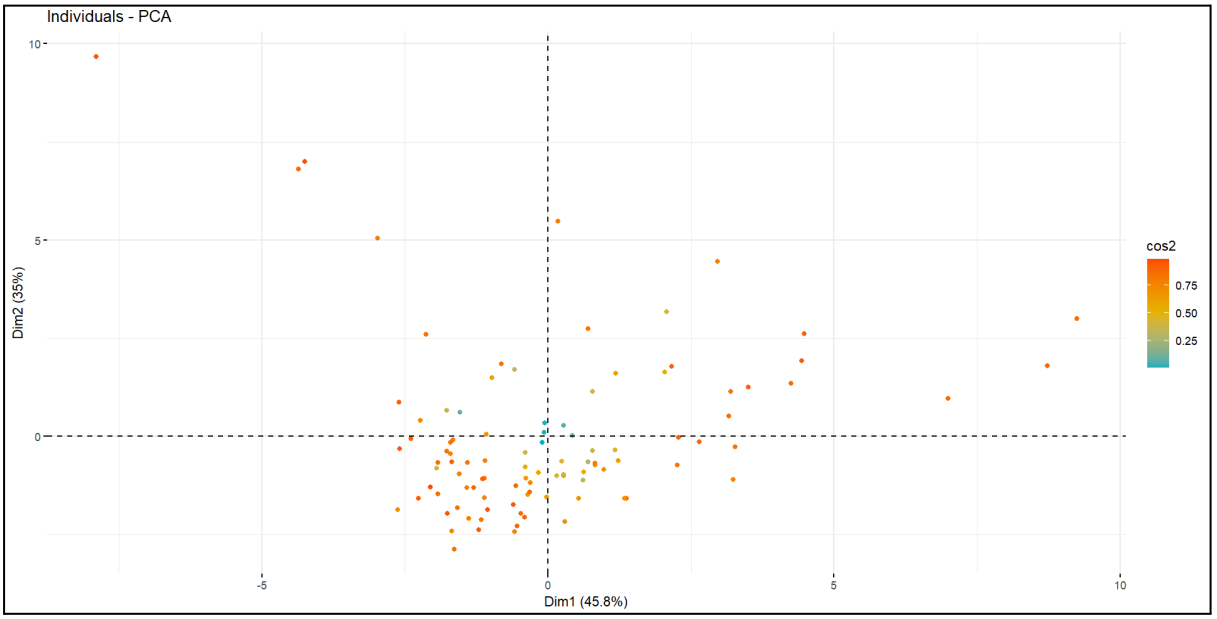


La variable Dépenses RSA brute est énormément corrélée au 4ème axe factoriel, et la variable Part des diplômés du supérieurs parmi les 25-34 ans est, elle, énormément corrélée au 5ème axe

Qualité de représentation

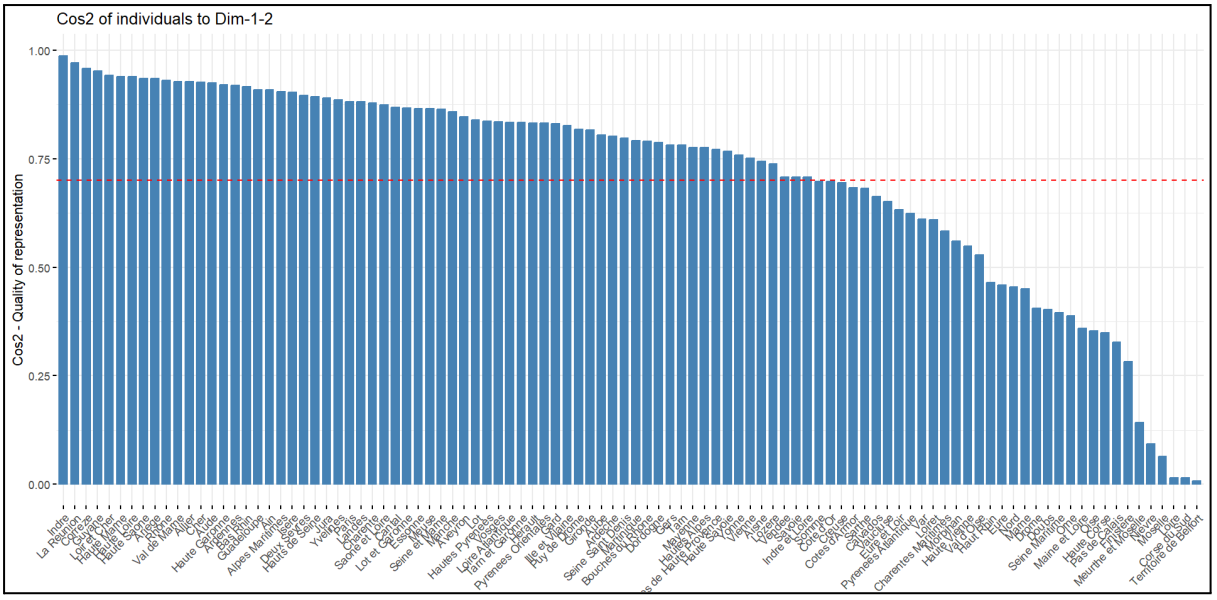
- **Cos2 des individus**

Il faut désormais interpréter les projections des individus. On sait qu'il ne faut pas faire confiance à une proximité entre deux individus sur le plan de projection pour en déduire une similarité entre les deux. On utilise donc un critère de qualité de représentation, celui du cos2. De façon arbitraire, on pourra interpréter la proximité entre deux individus sur le plan de projection si leurs qualités de représentation est > 0,7.
Pour les individus, on a :



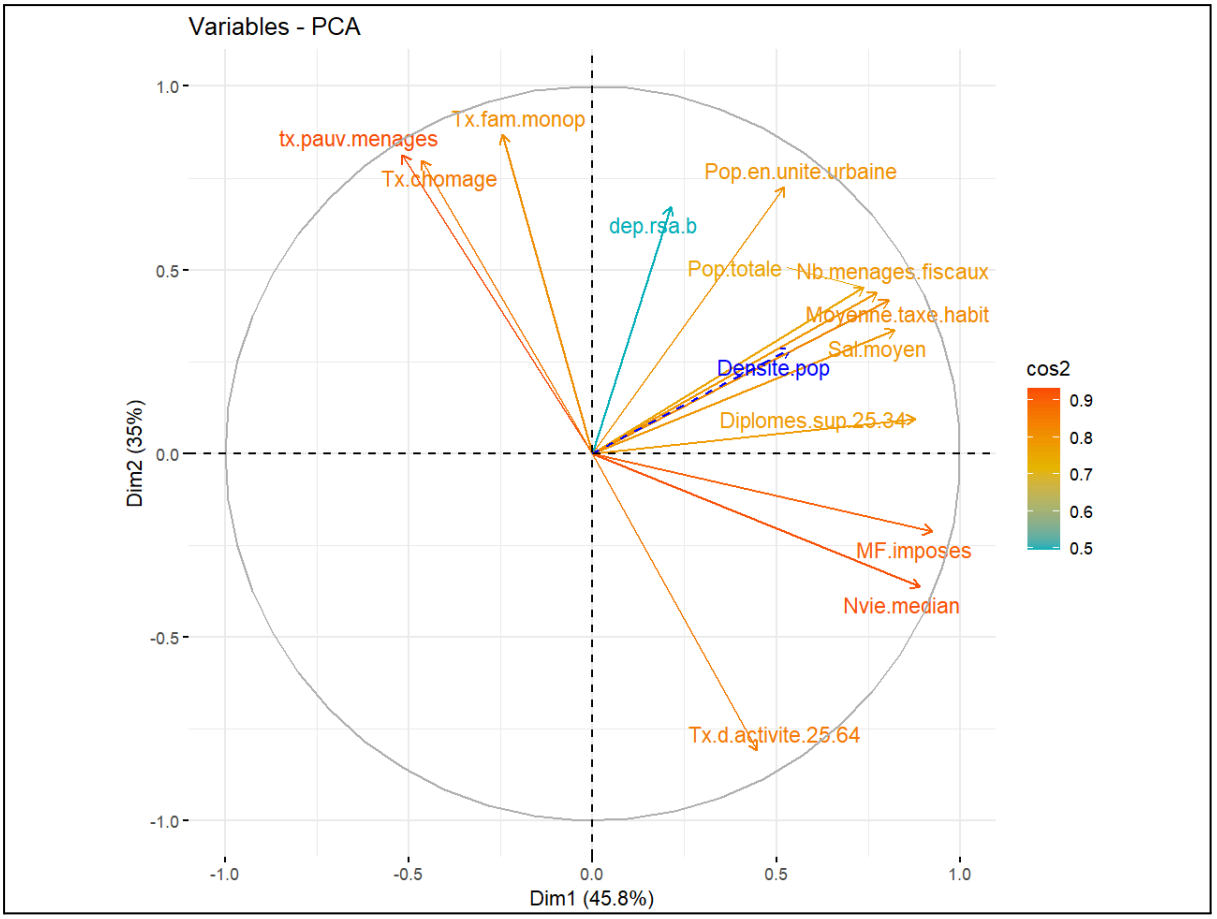
On a beaucoup d'individus, donc la représentation par gradient de couleurs est un peu plus lisible qu'accorder un point de taille proportionnelle à leurs qualité de représentation. La première conclusion est que tous nos individus ne sont pas propices à une interprétation de leurs proximités sur le plan de projection car on a peu de cos2 qui sont supérieurs à 0,7. En particulier, les individus qui sont projetés proche du centre du plan sont peu interprétables, ce qui n'est pas étonnant.

On va alors modéliser le cos2 sous la forme d'un histogramme afin de mieux l'analyser. La ligne de pointillés rouges représente un cos2 de 0,7, donc chaque individus supérieur à cette ligne sera interprétable :

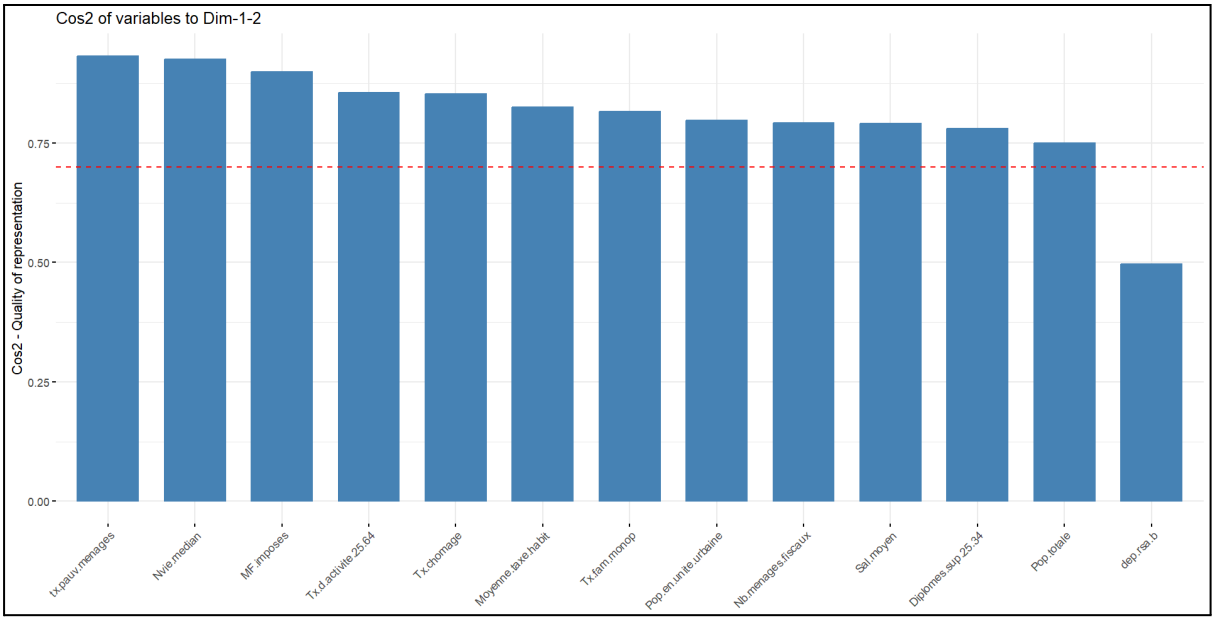


Finalement, deux tiers de nos individus sont sujets à l'interprétation d'une proximité sur le plan de projection.

- **Cos2 des variables**



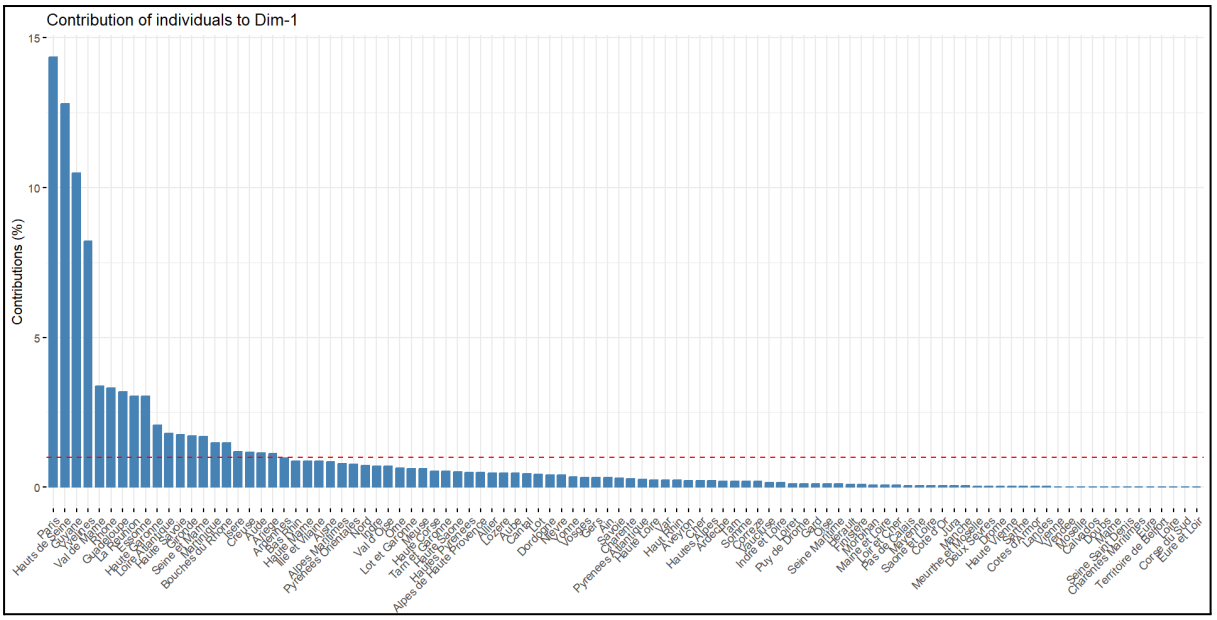
On a ici un gradient de couleurs en fonction de la qualité de représentation des variables. Plus la flèche est proche du bord du cercle, plus la variable est bien représentée. Avec le cos2, on a la même règle de comparaison, si le cos2 > 0,7 alors la variable est interprétable. La seule variable qu'on va avoir du mal à interpréter sera les dépenses brutes en RSA.
On propose également un histogramme, qui est plus lisible et précis :



Toutes les variables sont donc interprétables, à l'exception des dépenses RSA brutes.

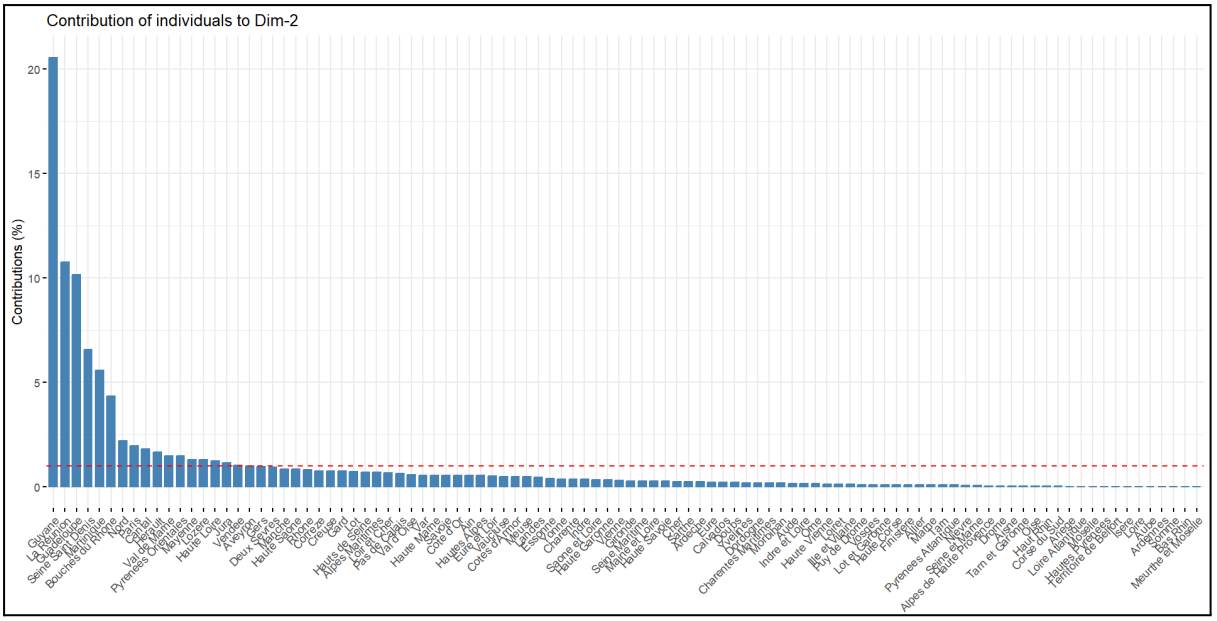
- **Contribution des individus aux axes**

On affiche désormais la contribution des individus aux deux premiers axes. Pour l'axe 1, on a :



Les individus qui contribuent particulièrement à l'axe 1 sont les départements "riches" d'Île-de-France, tels que Paris, les Hauts-de-Seine ou les Yvelines et les départements d'Outre-Mer, ainsi que certains autres départements plus peuplés que la moyenne tels que les Bouches-du-Rhône ou la Loire-Atlantique.
Ceci n'est pas étonnant lorsqu'on se rappelle que l'axe 1 va modéliser le niveau de richesse économique dans les départements. Les départements qui contribuent alors le plus à cet axe seront les départements qui ont tendance à avoir un niveau de richesse économique très élevé ou très faible.

Pour l'axe 2, on a :

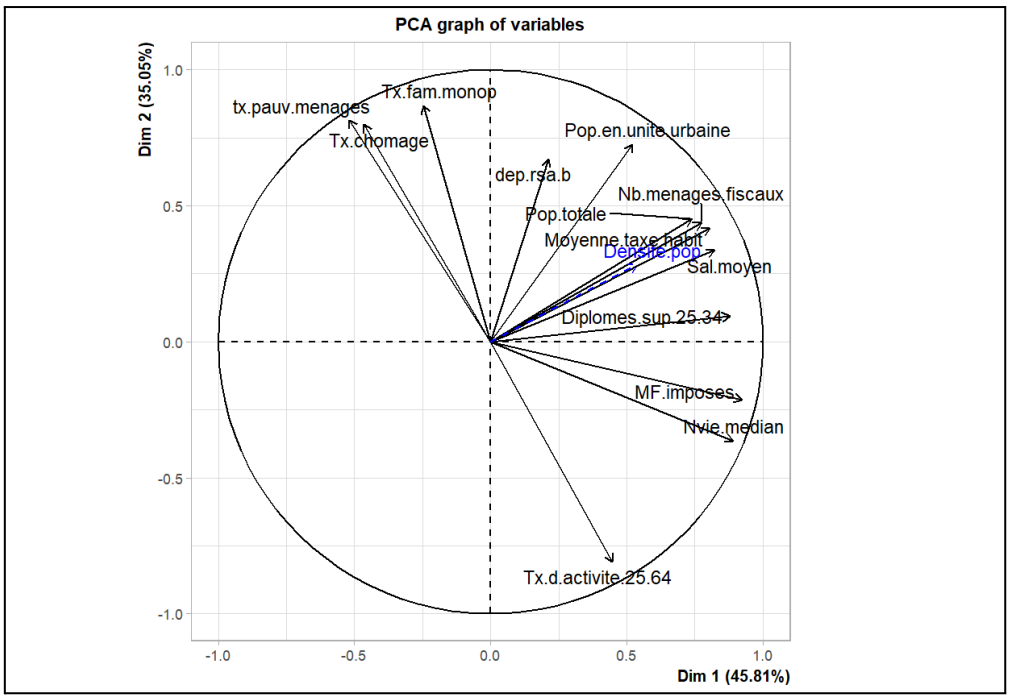


Ici, les individus qui participent le plus à l'axe 2 sont clairement les départements d'Outre-Mer, ainsi que certains départements d'Île-de-France, du Nord ainsi que certains départements connus pour un taux de précarité assez élevé associé à une densité de population très inférieure à la moyenne (le Cantal, la Lozère, l'Hérault …). Ce n'est pas étonnant, l'axe 2 représente la fragilité sociale. Malheureusement, on sait que les situations sociales dans les départements d'Outre-Mer (qui participent le plus à cet axe) sont extrêmement compliquées, avec des taux de pauvreté incomparables avec la France métropolitaine.

Interprétation des variables illustratives

- **Variables quantitatives illustratives**

Dans ce jeu de données, on avait qu'une seule variable quantitative supplémentaire : la densité de population au km2. On calcule donc la coordonnées de cette variable sur les axes, sa corrélation et son cos2.
On se rappelle qu'on avait ce plan de projection pour les variables :

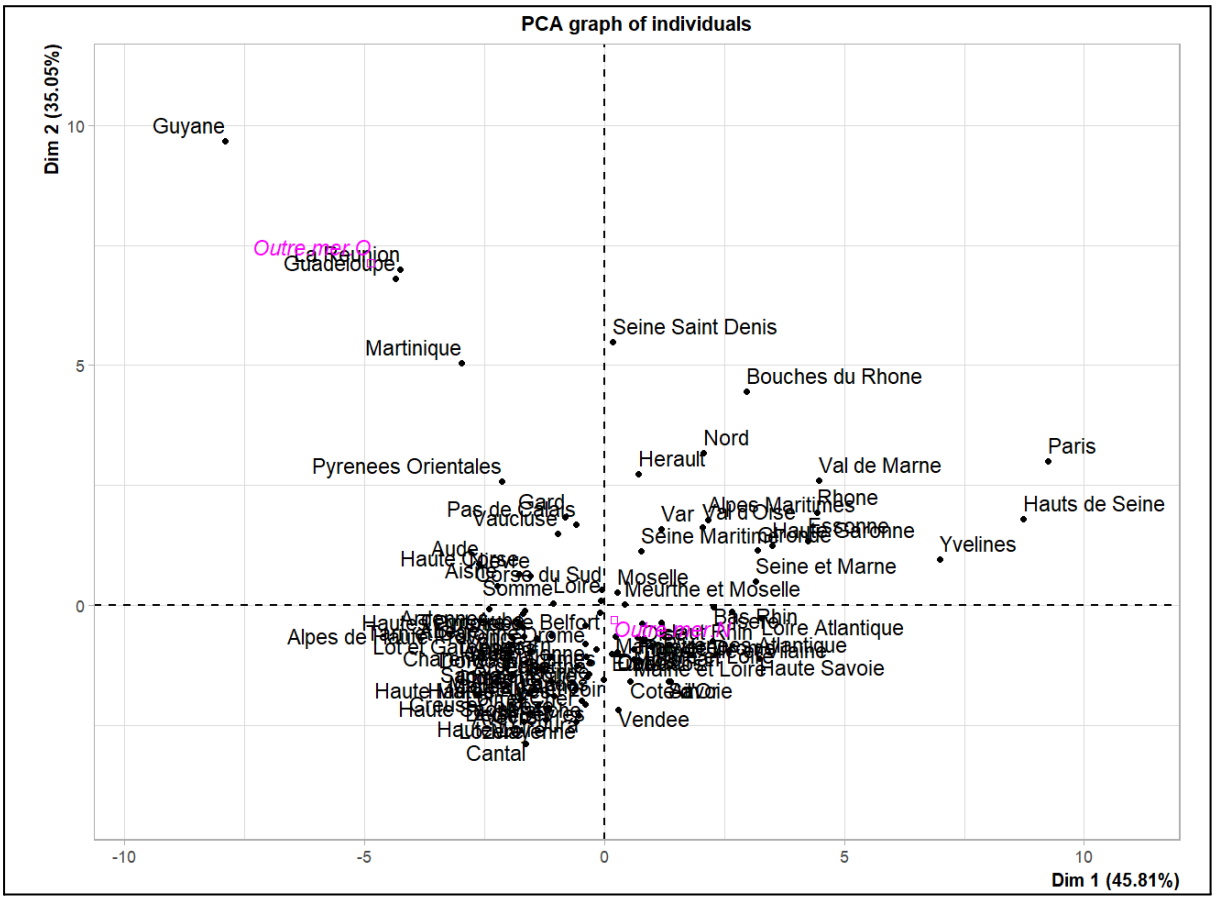



```
> pauv.acp1$quanti.sup
$coord
                  Dim.1     Dim.2
Densite.pop 0.5398988 0.2837317

$cor
                  Dim.1     Dim.2
Densite.pop 0.5398988 0.2837317

$cos2
                  Dim.1      Dim.2
Densite.pop 0.2914907 0.08050367
```

La variable quantitative supplémentaire est donc légèrement corrélée aux 2 axes, et est loin d'atteindre le cercle, donc sa qualité de représentation n'est pas bonne.

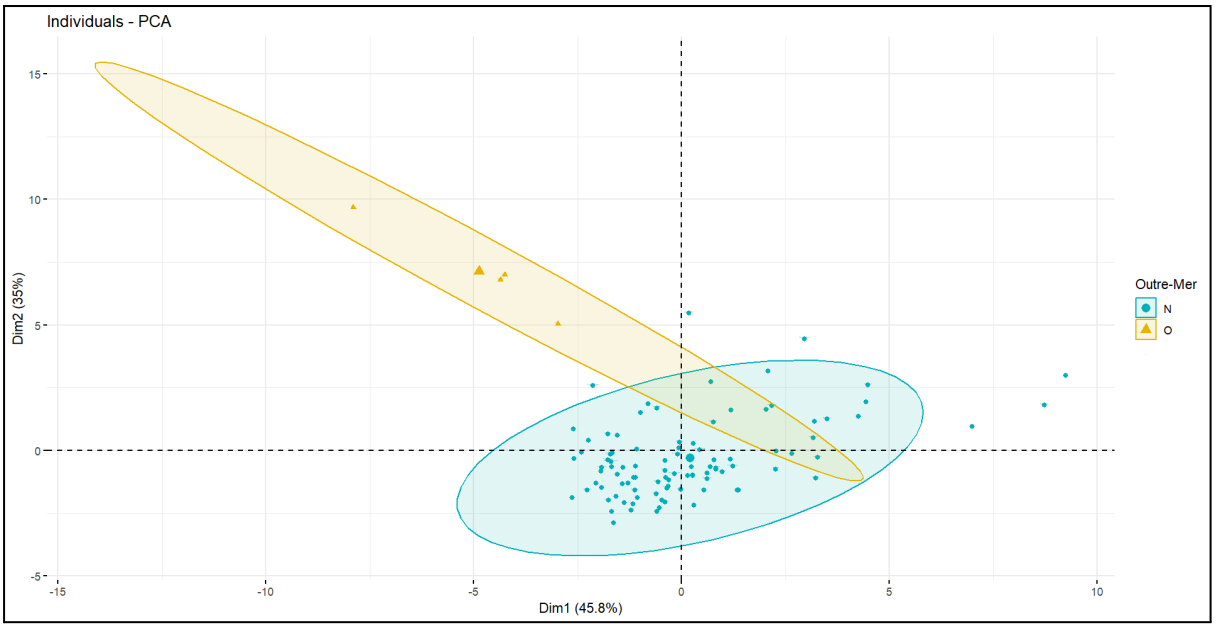
- **Variables qualitatives supplémentaires**

On avait ce plan de projection pour les individus :

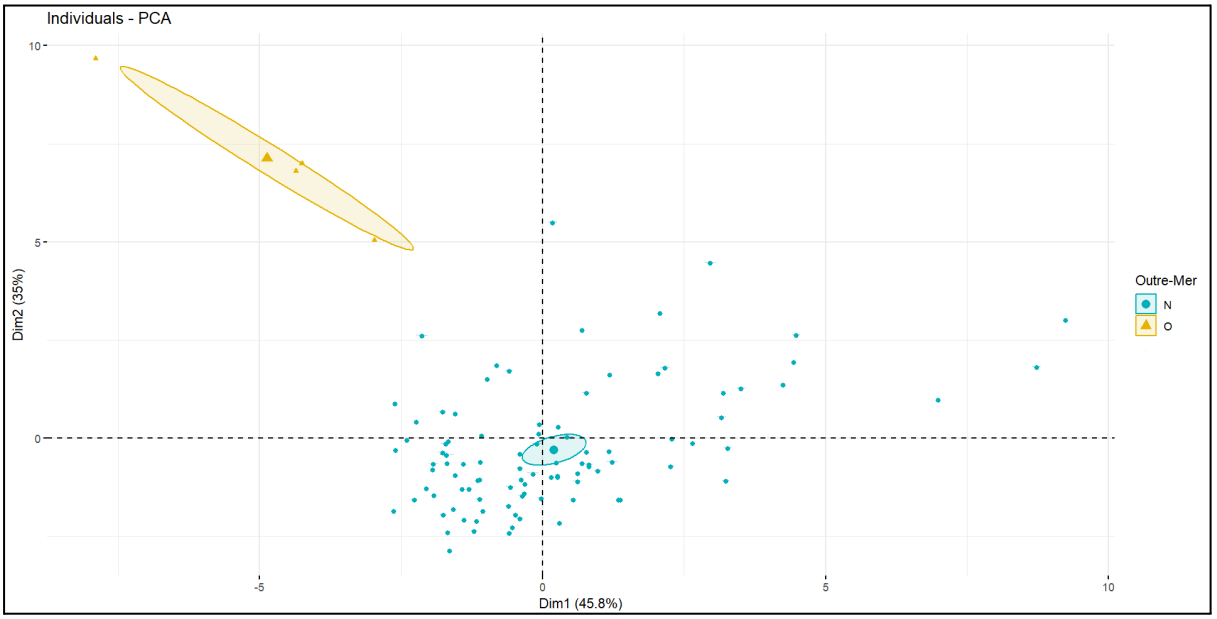


Avec les deux barycentres de la variable Outre-Mer en violet.

Pour pouvoir analyser un peu plus l'importance de cette variable, on utilise une ellipse de type "concentration" d'abord :



Puis avec une ellipse de type “confidence”, qui incluera 95% de mes individus si les données sont distribuées selon une loi normale :



Les données ne sont évidemment pas distribuées selon une loi normale.

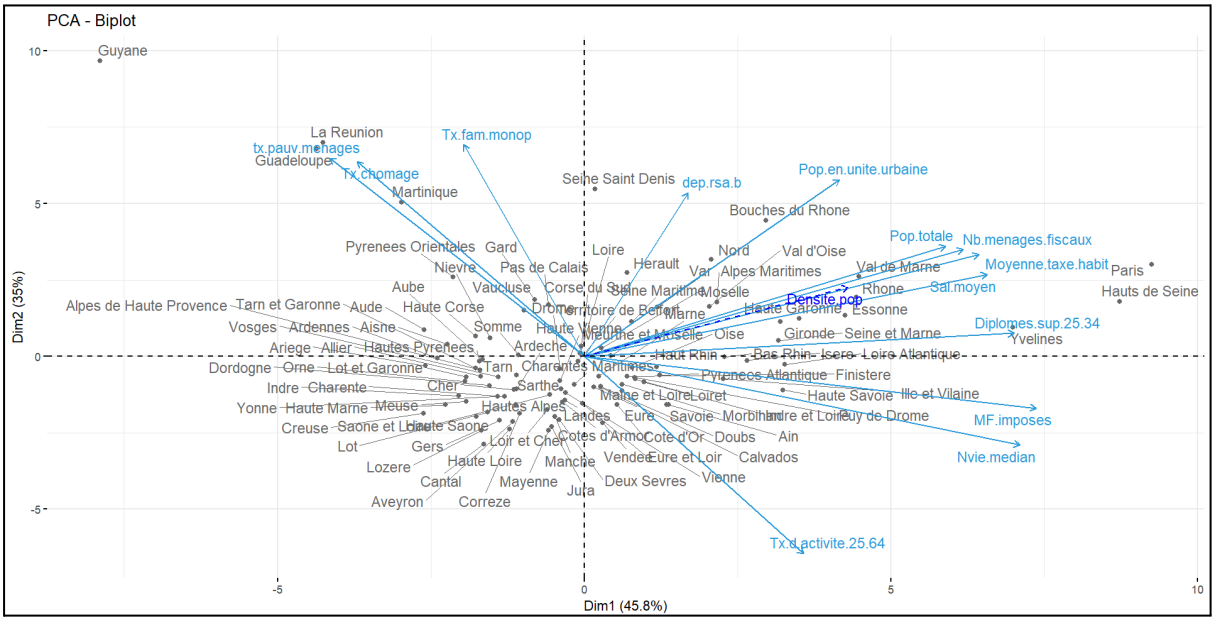
On remarque que la différence entre les deux groupes sur le plan de projection est très marquée, mais est-ce que ces modalités sont significatives ? Pour le savoir, on va utiliser la V.Test (en valeur absolue) qu’on va comparer à 1,96 :

```
> pauv.acp1$quali.sup
$coord
                Dim.1      Dim.2       Dim.3        Dim.4       Dim.5
Outre.mer.N  0.2029369 -0.2971413 -0.04755148  0.001860202 -0.01976087
Outre.mer.O -4.8704859  7.1313922  1.14123547 -0.044644852  0.47426089

$cos2
               Dim.1     Dim.2      Dim.3        Dim.4       Dim.5
Outre.mer.N 0.308586 0.6615767 0.01694268 2.592832e-05 0.002925945
Outre.mer.O 0.308586 0.6615767 0.01694268 2.592832e-05 0.002925945

$v.test
                Dim.1     Dim.2     Dim.3      Dim.4     Dim.5
Outre.mer.N  4.053449 -6.785357 -2.590244  0.1216166 -1.643917
Outre.mer.O -4.053449  6.785357  2.590244 -0.1216166  1.643917

$dist
Outre.mer.N Outre.mer.O
  0.3653196   8.7676704

$eta2
              Dim.1     Dim.2      Dim.3        Dim.4     Dim.5
Outre.mer 0.1659641 0.4650613 0.06777134 0.0001493999 0.0272976
```

On conclut alors que les modalités de la variable Outre-Mer sont significatives pour les deux premiers axes. On peut également interpréter le coefficient eta2, qui illustre la corrélation de la variable supplémentaire avec l'axe. Outre-Mer est très corrélée avec la 2ème composante principale.

Conclusion

On termine avec biplot pour explorer les relations entre individus et variables :



On remarque que les départements d'Outre-Mer ont tendance à prendre des valeurs plus élevées que la moyenne pour des variables telles que le taux de pauvreté, le taux de chômage ou le taux de familles monoparentales, et à l'inverse des valeurs moins élevées que la moyenne pour le taux d'activité. Ce sont les variables qui étaient extrêmement corrélées à l'axe 2.
De la même manière, on remarque que les départements d'Île-de-France tels que Paris, les Hauts-de-Seine ou les Yvelines ont tendance à prendre des valeurs plus élevées que la moyenne pour la part des diplômés du supérieur parmi les 25-34 ans, le salaire mensuel moyen ou encore la moyenne de la taxe d'habitation.

Pour conclure, l'ACP réalisée a mis en évidence l'existence de deux groupes parmi nos individus : les départements d'Outre-Mer, et les départements de France métropolitaine. Le premier plan factoriel récupère 80% de l'inertie totale du jeu de données, et les deux premiers axes factoriels sont les axes qui sont intéressants à interpréter.

Ainsi, on a abouti à deux composantes principales qui vont modéliser le niveau de richesse, la prospérité économique, la qualité de vie pour la première et la fragilité sociale pour la deuxième.

Sources

- Taux de pauvreté par ménages en France en 2020 : https://www.insee.fr/fr/statistiques/6692414?sommaire=6692394#tableau-figure1_radio1
- Niveau de vie médian en 2021 : https://www.insee.fr/fr/statistiques/2012717
- Taux de chômage en France en 2020 : créé à partir de https://www.insee.fr/fr/statistiques/2012804#tableau-TCRD_025_tab1_departements (données accessibles sur le fichier .csv téléchargeable)
- Part des diplômés du supérieur parmi les 25-34 ans en 2021 : https://www.insee.fr/fr/statistiques/5020064?sommaire=5040030#tableau-figure1_radio1
- Salaire mensuel net moyen en 2020 : https://www.journaldunet.com/business/salaire/classement/departements/salaires
- Dépenses RSA brutes en 2020 : https://data.drees.solidarites-sante.gouv.fr/explore/dataset/bases-de-donnees-brutes-de-l-enquete-aide-sociale-volet-depenses/information/ puis ouvrir Base Dépenses 2020 et lire colonne RSA1_A7 (Allocation RSA)
- Nombre de ménages fiscaux en 2021 : https://www.insee.fr/fr/statistiques/2012717
- Part des ménages fiscaux imposés en 2021 : https://www.insee.fr/fr/statistiques/2012717
- Population totale en 2020 : https://www.insee.fr/fr/statistiques/6683015?sommaire=6683037
- Taxe d'habitation moyenne par résidence principale : https://www.impots.gouv.fr/dgfip-statistiques-les-impots-locaux-des-particuliers-en-2020
- Taux de famille monoparentales en 2021 : créé à partir de https://www.observatoire-des-territoires.gouv.fr/nombre-de-familles-monoparentales et de https://www.insee.fr/fr/statistiques/2012696#tableau-TCRD_007_tab1_departements
- Population vivant dans une unité urbaine en 2017 : https://www.insee.fr/fr/statistiques/4806684#tableau-figure2
- Taux d'activité chez les 25-54 ans en 2021 : https://www.insee.fr/fr/statistiques/2012710#tableau-TCRD_015_tab1_departements
- Densité de population au km2 en 2021 : https://www.insee.fr/fr/statistiques/7666835?sommaire=7666953#tableau-figure3